Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

深度生成模型

博士研究生　周瑾洁

2022年1月9日

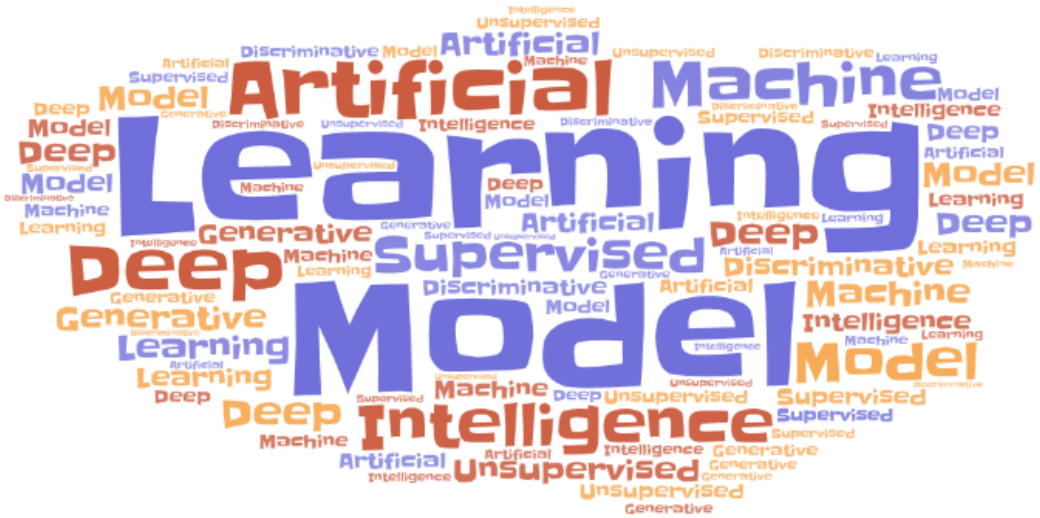
内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

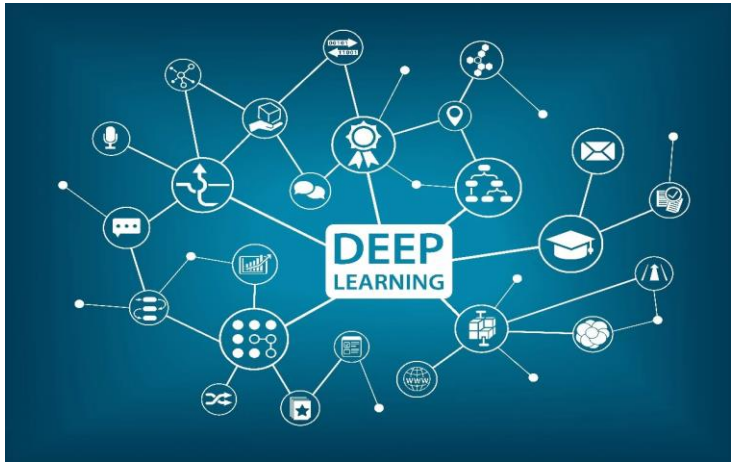
背景简介

- 预期收获
  - 1.了解机器学习中有监督学习和无监督学习的区分概念
  - 2.了解判别模型和生成模型的区分概念
  - 3.理解去噪自动编码和去噪匹配打分网络的底层意义
  - 4.理解噪声条件打分网络的工作原理

背景简介

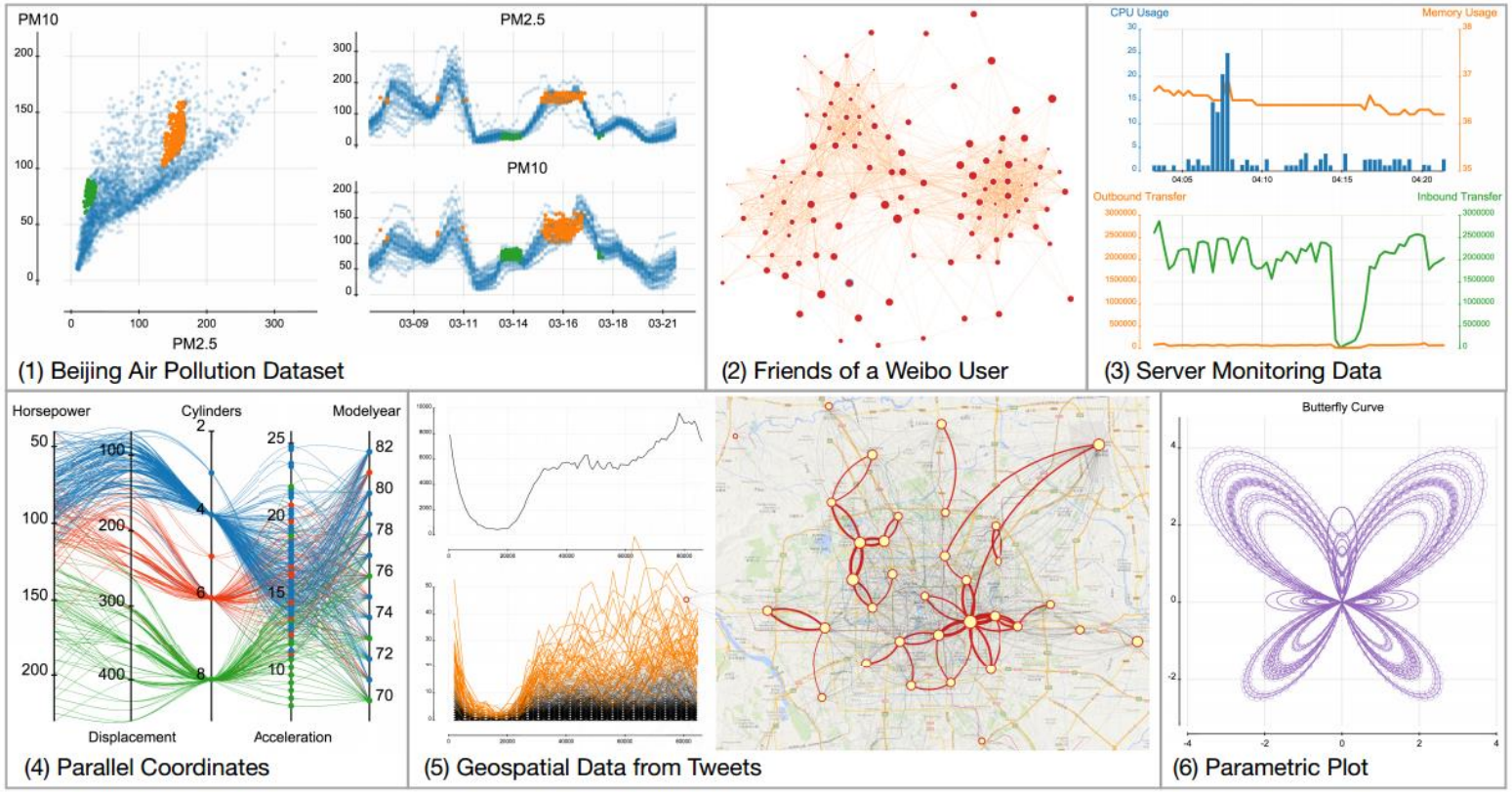


- 不少人对这些高频词汇的含义及其背后的关系总是似懂非懂、一知半解



背景简介

- 在大数据背景下，高维数据灾难是机器学习、数据压缩、模式识别、可视化等领域面临的关键问题之一
- 在高维数据下，机器学习会存在流形假设和低数据密度区域两大问题



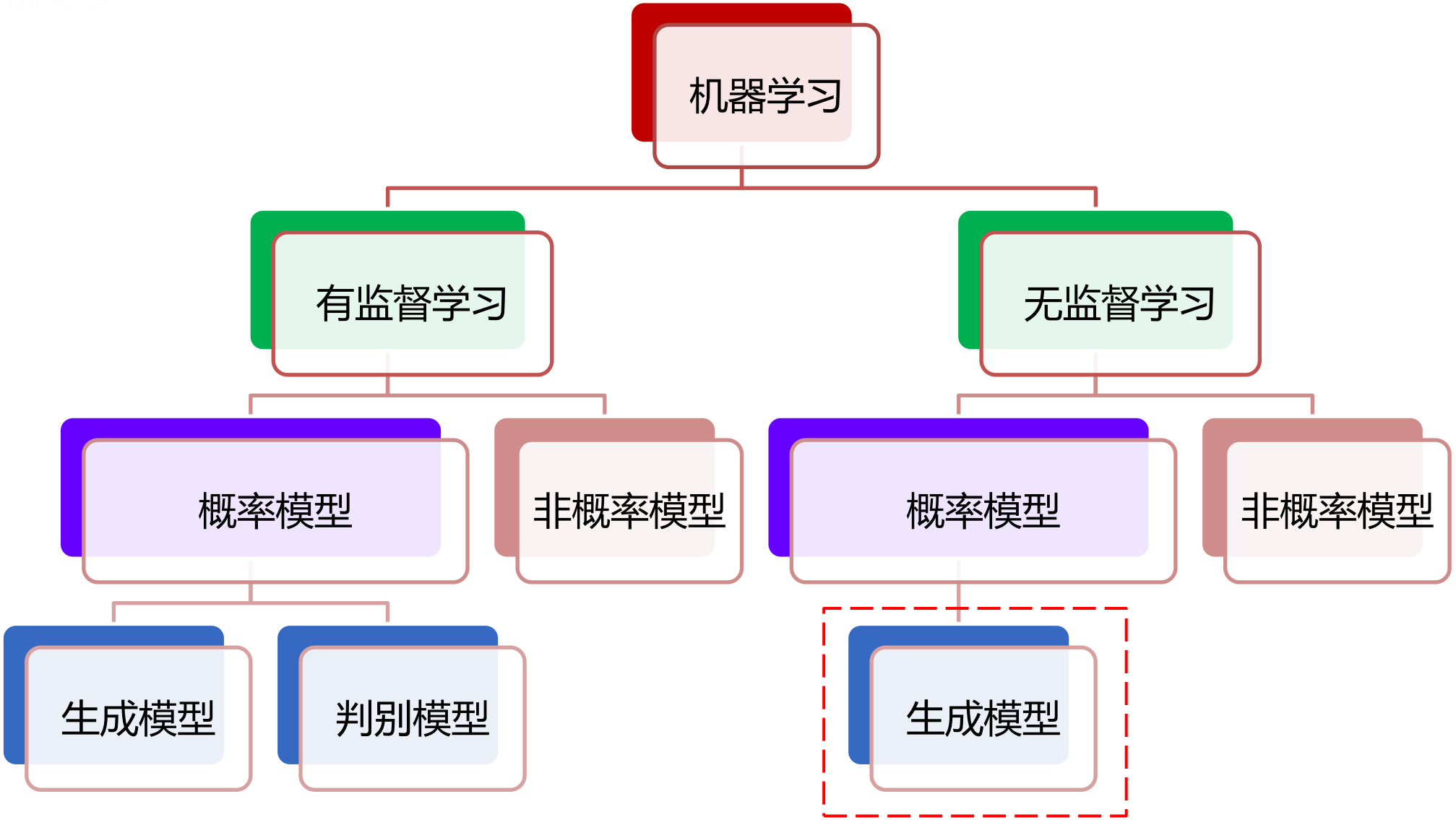
基本概念

基本概念

- 人工智能
  - 研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门技术科学
- 机器学习
  - 用算法解析数据，不断学习，对世界中发生的事做出判断和预测的一项技术
- 深度学习
  - 用于建立、模拟人脑进行分析学习的神经网络，并模仿人脑的机制来解释数据的一种机器学习技术

学习（百度百科）:是指通过阅读、听讲、思考、研究、实践等途径获得知识的过程。

基本概念



基本概念

- 有监督学习和无监督学习
  - 是否为有监督需要看输入的数据是否含有标签（Label）
  - 数据含有标签，为有监督；不含有标签，为无监督
- 半监督学习
  - 综合利用有类标签的数据和没有类标签的数据，来生成合适的分类函数。利用少量标注样本和大量未标注样本进行机器学习，从概率学习角度可理解为研究如何利用训练样本的输入边缘概率 $P(x)$和条件输出概率$P\ (y|x)$的联系设计具有良好性能的分类器

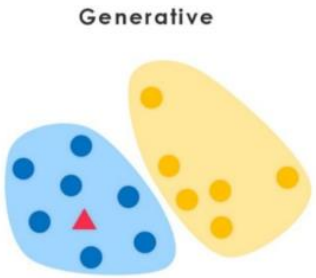
基本概念

- 生成模型和判别模型（有监督）
  - 生成模型：由数据学习联合概率分布$\boldsymbol{p}(\boldsymbol{x},\boldsymbol{y})$，然后求出条件概率分布作为预测的模型。即给定$x$产生出$y$的生成关系
    - 典型比如：朴素贝叶斯方法、隐马尔科夫方法等
  - 判别模型：由数据直接学习决策函数或者条件概率分布$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$作为预测模型。给定x应该预测什么样的输出y
    - 典型比如：**KNN**、感知机、决策树、逻辑斯蒂回归、最大熵、**SVM**、提升方法、条件随机场、等

基本概念

生成模型和判别模型各自特点：

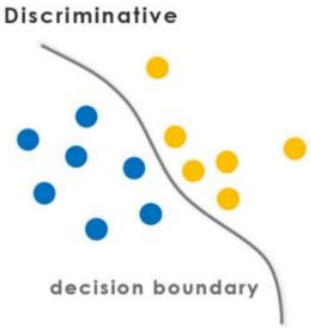
- 生成模型



  - 优点：信息丰富，单类问题灵活性强，增量学习，合成缺失数据
  - 缺点：学习过程复杂，为分布牺牲分类性能

由生成模型可以得到判别模型，但由判别模型得不到生成模型。

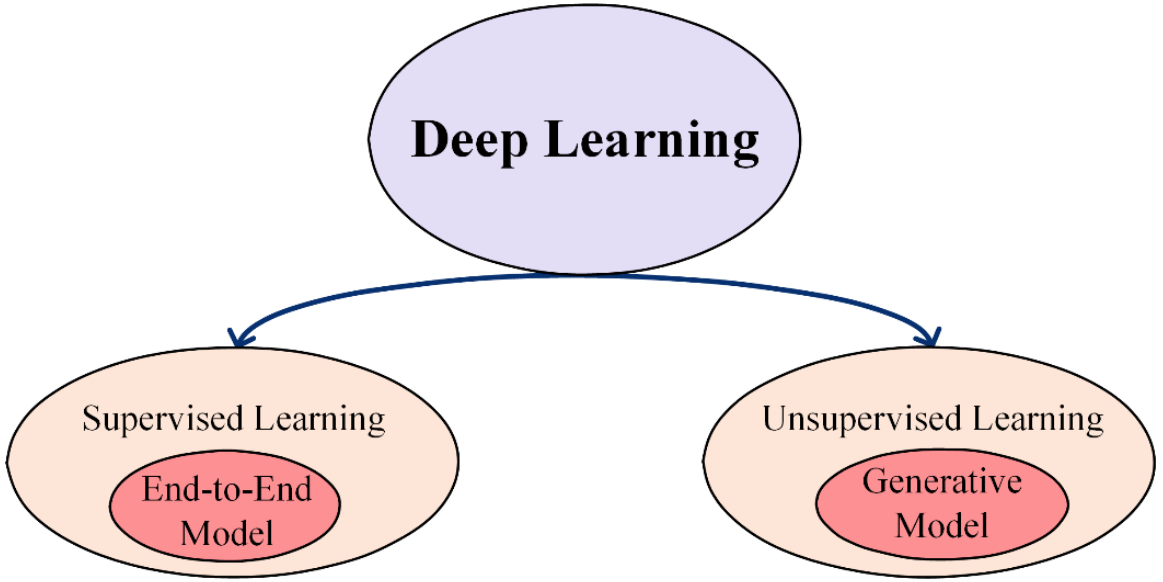
- 判别模型



  - 优点：类间差异清晰，分类边界灵活，学习简单，性能较好
  - 缺点：不能反应数据特征、需要全部数据进行学习

基本概念

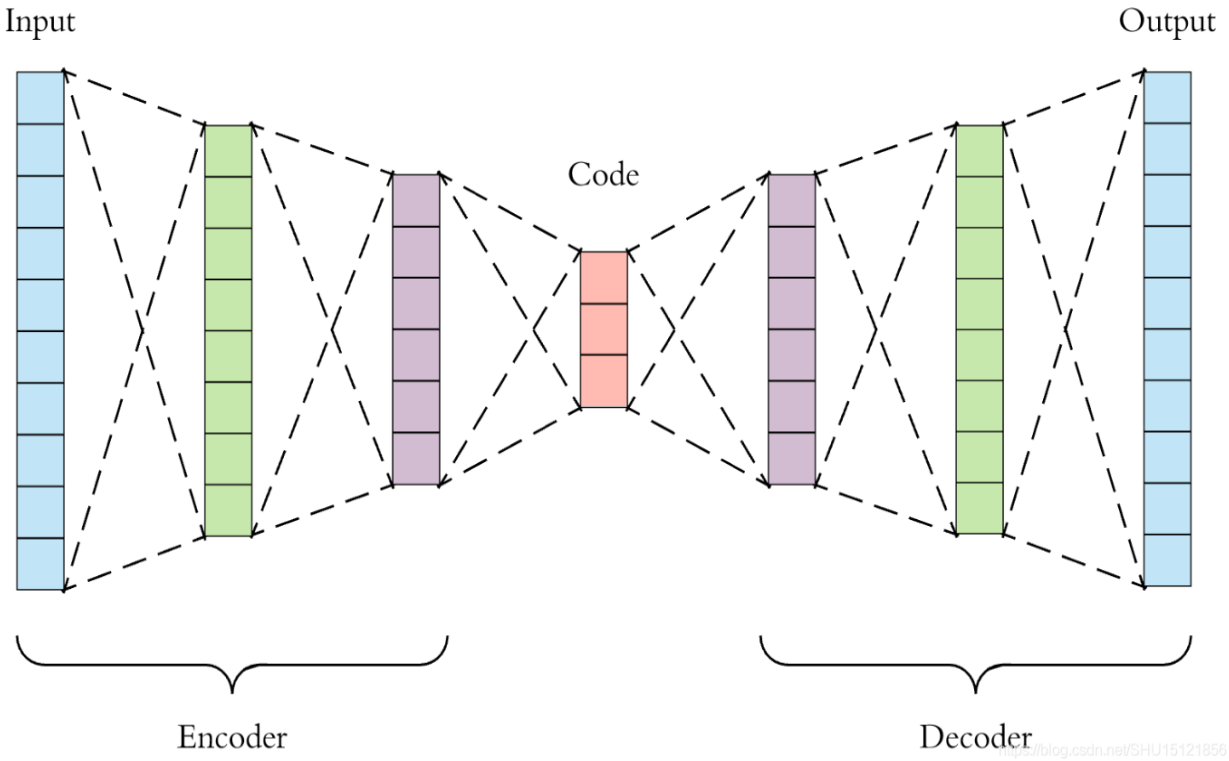
- 生成模型（无监督）
  - 由数据学习联合概率分布$p(x, y)$，然后求出条件概率分布作为预测的模型。即给定x产生出y的生成关系
  - 典型比如：**DAE、VAE、GAN、pixelCNN、Glow等**



Generative model

基本概念

- 自编码器AutoEncoder
  - 编码：将原始表示编码成隐层表示
  - 隐层：比原始表示的维度更低，以得到更稠密更有意义的表示
  - 解码：将隐层表示还原(Reconstruction)成原始表示形式
  - 训练目标：最小化重构误差
  - 缺点：容易过拟合



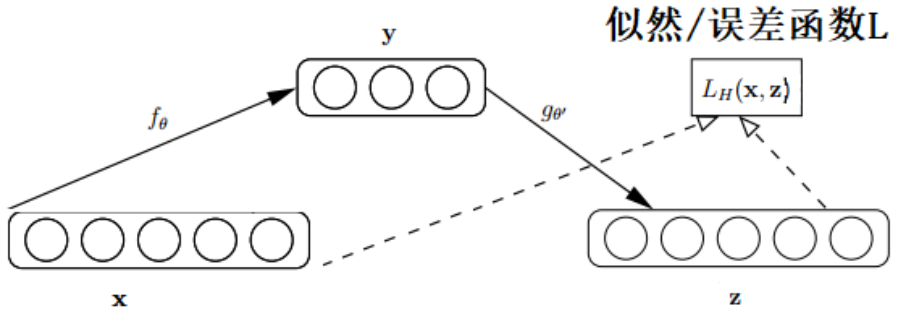
AE实际上就是一个特殊的全连接网络：

① 输入和输出的维度一样

② 网络中间有一个neck，可以取出其中的隐层表示

基本概念

- 自编码器**AutoEncoder**
  - 初始**input**（设为x）经过加权（W、b)、映射（**Sigmoid**）之后得到y，再对y反向加权映射回来成为z
  - 通过反复迭代训练两组（W、b），使得误差函数最小，即尽可能保证z近似于x，即完美重构了x



学了单词：Easy

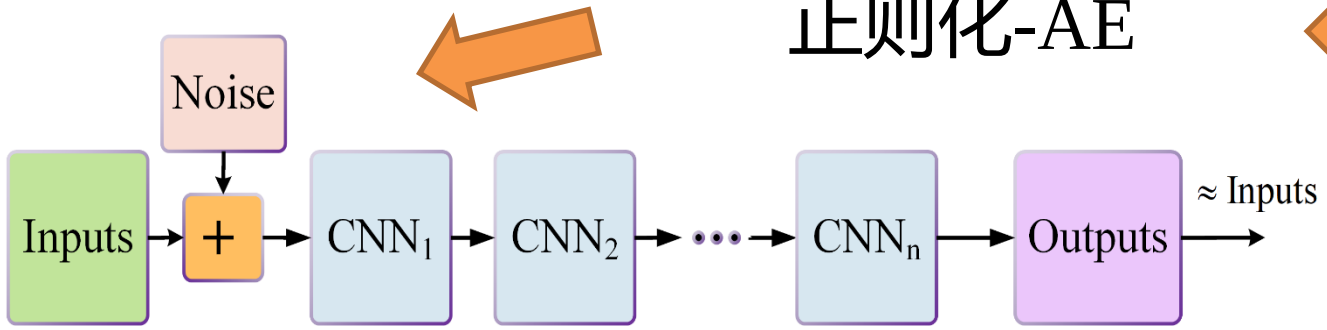
简单的英语?

是easy啊！

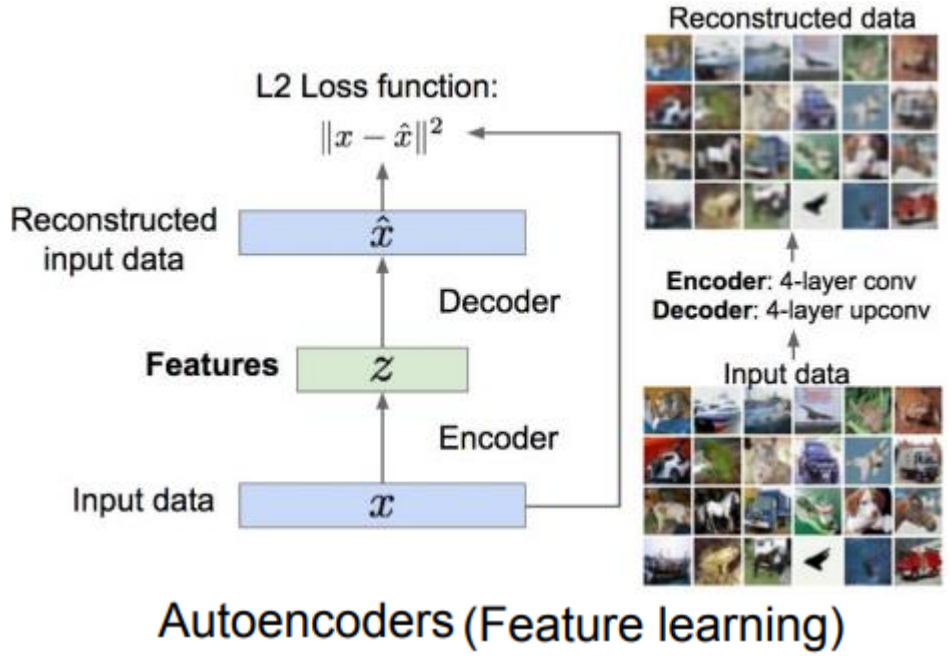
尽管少了首字母大写

基本概念

- 去噪自编码器 **AutoEncoder**
  - 加入随机噪声
  - 结合正则化思想

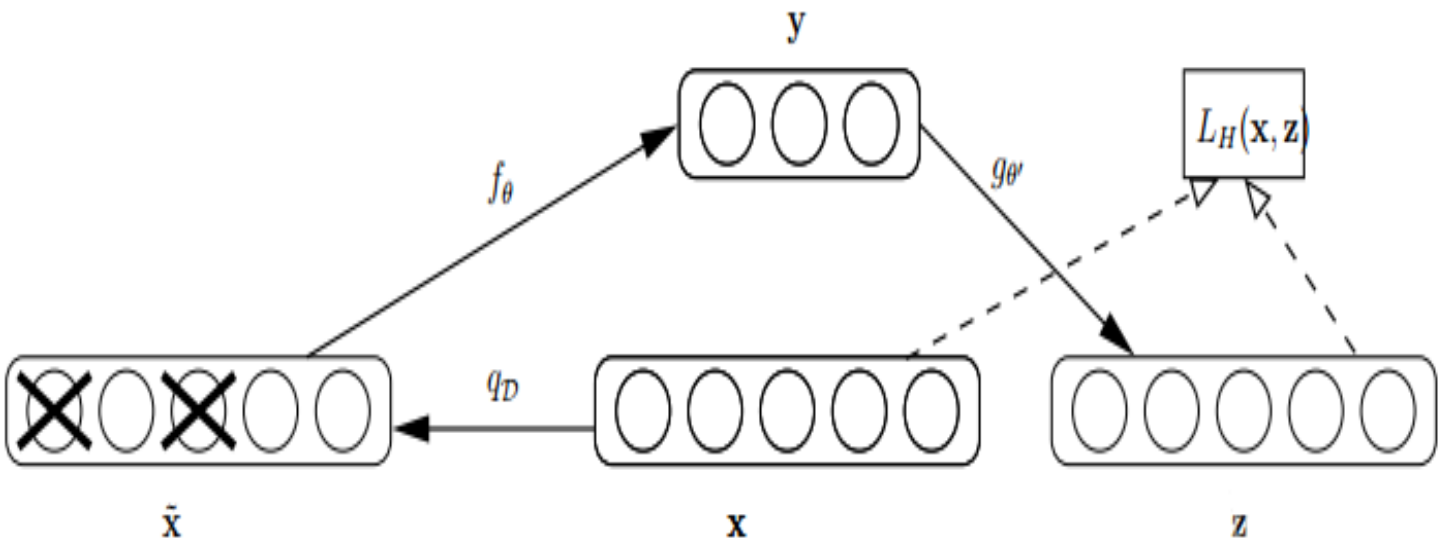


$$L_{DAE}(A) = E_{x,\,\mu}[\|A_{\sigma}(x+\mu) - x\|^2]$$



基本概念

- 去噪自编码器**Denoising AutoEncoder**
  - 去擦除原始**input**矩阵，即每个值都随机置0，部分数据的部分特征丢失
  - 以这丢失的数据x'去计算y，计算z，并将z与原始x做误差迭代，网络学习了破损数据



基本概念

- 去噪自编码器Denoising AutoEncoder
  - 通过与非破损数据训练的对比，破损数据训练出来的Weight噪声比较小。
  - 破损数据一定程度上减轻了训练数据与测试数据的代沟。

学了单词：worker

学习过程（编码）

不就是work+er
已记忆worker词根

学习过程（破损）

重建过程

算法原理

算法原理

T	实现图像恢复
I	受损图像
P	1.输入加噪声 2.构建网络架构 3.学习DAE先验信息 4.计算数据梯度，计算先验梯度 5.恢复图像
O	恢复后的图像

P	原始AE过拟合
C	输入受损
D	恢复图像时，梯度比较难求
L	2018 CCF B类会议

算法原理

- 去噪自编码先验**Denoising Autoencoders Prior（DAEP）**
  - 损失函数 $L_{DAE}(A) = E_{x\sim p,\eta\sim g_\sigma}[\|A(x+\eta)-x\|^2]$
  - 网络输出$g_{\sigma_\eta}(\eta)$与真实数据密度$p(u)$的关系

$$A_{\sigma_\eta}(x) = \frac{\int (x-\eta) g_{\sigma_\eta}(\eta) p(x-\eta) d\eta}{\int g_{\sigma_\eta}(\eta) p(x-\eta) d\eta} = x - \frac{\int g_{\sigma_\eta}(\eta) p(x-\eta)\eta d\eta}{\int g_{\sigma_\eta}(\eta) p(x-\eta) d\eta}$$

$$= x + \sigma_\eta^2 \nabla \int g_{\sigma_\eta}(\eta) p(x-\eta) d\eta = x + \sigma_\eta^2 \nabla \log[\, g_{\sigma_\eta} * p](x)$$

$$\left\| A_{\sigma_\eta}(x) - x \right\|^2 = \left\| \sigma_\eta^2 \nabla \log[\, g_{\sigma_\eta} * p](x) \right\|^2$$

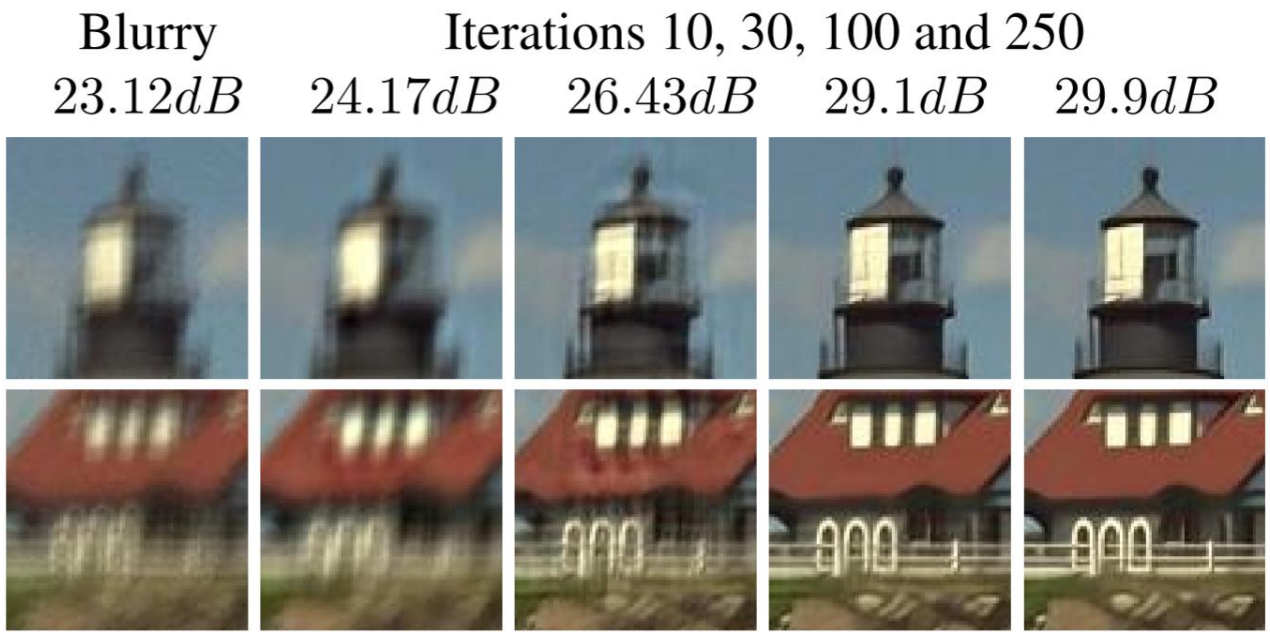
先验信息

算法原理

- 去噪自编码先验**Denoising Autoencoders Prior（DAEP）**

$$\left\| A_{\sigma_\eta}(x) - x \right\|^2 = \left\| \sigma_\eta^2 \nabla \log[\, g_{\sigma_\eta} * p](x) \right\|^2$$

先验信息辅助图像去模糊



算法原理

- 深度均值偏移先验 **Deep Mean-shift Priors（DMSP）**
  - 构造了一个与自然图像的高斯平滑概率分布的对数相对应的先验

$$prior(x) = \log \int g_{\sigma_\eta}(\eta)\, p(x+\eta)\, \mathrm{d}\,\eta$$

  - 随后通过DAE学习等式中的先验梯度$L_{DAE}(A) = E_{x\sim p,\eta\sim g_\sigma}[\left\| A_{\sigma_\eta}(x+\eta) - x \right\|^2]$

$$A_{\sigma_\eta}(x) = x - \frac{\int g_{\sigma_\eta}(\eta) q(x-\eta)\eta d\eta}{\int g_{\sigma_\eta}(\eta) q(x-\eta) d\eta} = x + \sigma_\eta^2 \nabla \log \int g_{\sigma_\eta}(\eta) q(x-\eta) d\eta$$

  - 先验的梯度对应于自然图像分布上的均值偏移向量

$$\nabla \mathrm{prior}(x) = \nabla \log \int g_{\sigma_\eta}(n) p(x+\eta) dn = \left[A_{\sigma_\eta}(x) - x \right] / \sigma^2$$

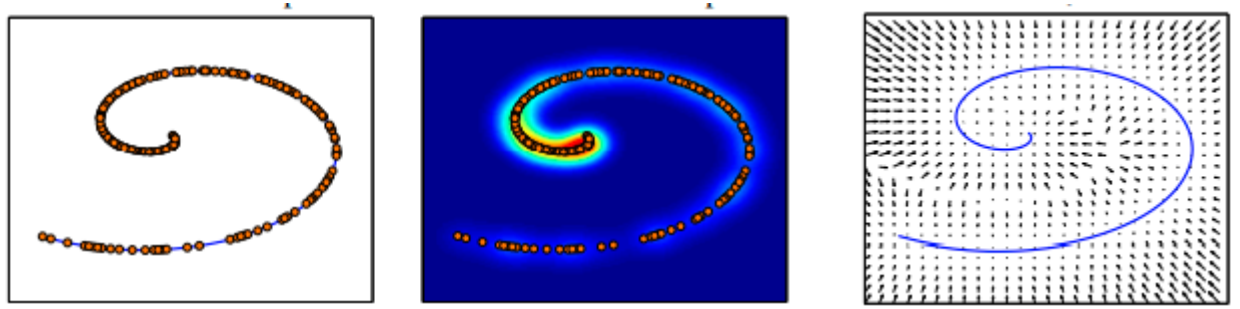
算法原理

- 深度均值偏移先验 **Deep Mean-shift Priors（DMSP）**
  - 先验的梯度对应于自然图像分布上的均值偏移向量

$$\nabla \text{prior}(x) = \nabla \log \int g_{\sigma_\eta}(n) p(x+\eta) dn = \left[A_{\sigma_\eta}(x) - x \right] / \sigma^2$$

  - 使用2D螺旋密度可视化DAE

(a)流型螺旋和观察到样本　(b)观察样本的平滑密度　(c) DAE学习的均值偏移向量



算法原理

- **Score Matching(SM)和 Denosing Score Matching(DSM)**
  - **Score Function 是概率的log值的梯度(the gradient of the log density)**
  - 我们对人工扰动的数据对 $(x，\ \tilde{x})$ 进行学习，我们可以定义我们的优化目标为

$$J_{DSM_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x},x)}\left[\frac{1}{2}||\Psi(\tilde{x};\theta) - \frac{\partial log q_\sigma(\tilde{x}|x)}{\partial \tilde{x}}||^2\right.$$

  - 其中 $q_\sigma(\tilde{x},x) = q_\sigma(\tilde{x}|x)q_0(x) = \frac{1}{(2\pi^{d/2}\sigma^d)} e^{-\frac{1}{2\sigma^2}||\tilde{x}-x||^2} \frac{1}{n}\sum_{i=1}^{n}\delta(||x-x^{(i)}||)$
  - 是一个噪声数据和真实数据对的联合概率分布
  - 则DSM的训练 $\frac{\partial log q_\sigma(\tilde{x}|x)}{\partial \tilde{x}} = \frac{1}{\sigma^2}(x-\tilde{x})$

算法原理

DAEP

$$L_{DAE}(A) = E_{x\sim p,\ \eta\sim g_\sigma}[\|A(x+\eta)-x\|^2]$$

DMSP

$$\nabla prior(x) = \nabla \log \int g_\sigma(\eta)\, p(x+\eta)\, d\eta$$

$$= [(A_\sigma(x)-x)]/\sigma^2$$

DSM

$$L_{DSM}(S) = E_{p_\sigma}[\|S(x)-\nabla \log p_\sigma(x)\|^2]$$

$$S(x) = \frac{A(x)-x}{\sigma^2}$$

$$S(x) \to \nabla \log p_\sigma(x)$$

NCSN

DAE与DSM在$S(x) = \frac{A(x)-x}{\sigma^2}$是等价的并且不依赖于$A$和$S$

算法原理

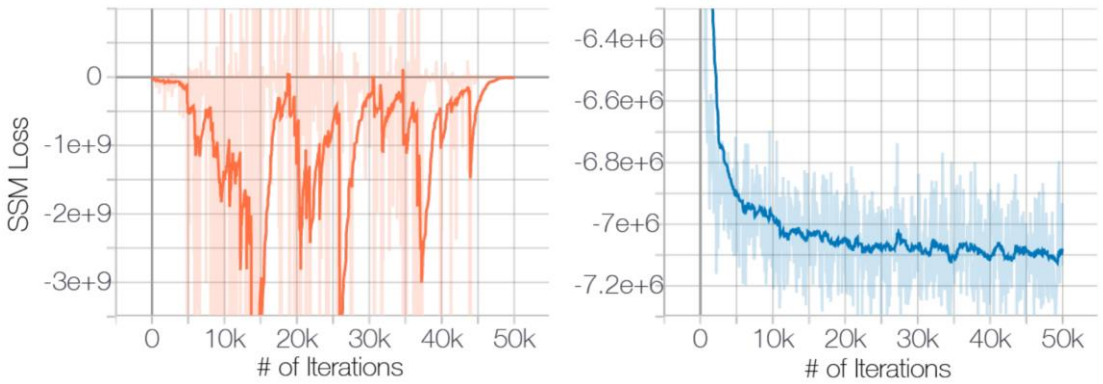
算法原理

T	实现高质量图像生成和填充
I	原始的高斯噪声
P	1.利用DSM构建打分网络 2.利用打分网络学习数据分布的梯度 3.朗之万模拟退火进行采样生成图像
O	高质量图像

P	原始DAEP等算法计算成本较高，机器学习中的流行假设和低数据密度区域
C	合适的噪声和退火噪声等级
D	如何应于不同的图像任务
L	NeurIPS 2019

算法原理

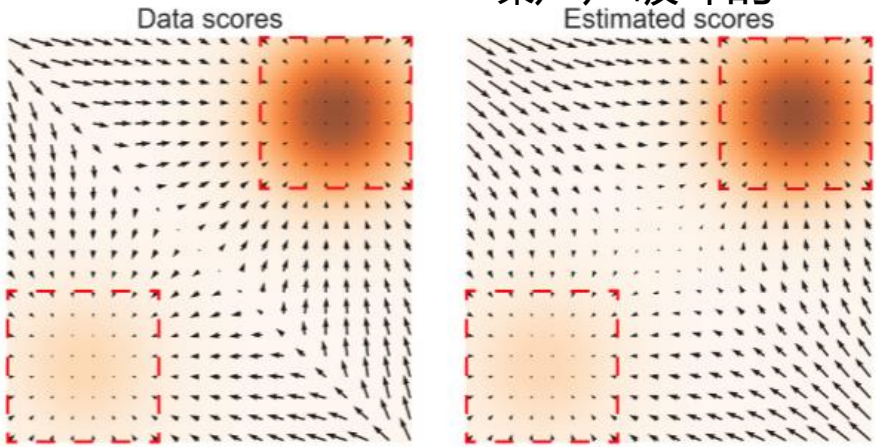
- 去噪条件打分网络（NCSN）
  - 流行假设



没有加噪声的**SMloss**

被$N(0,0.0001)$ 分布的噪声声破坏的**SMloss**

  - 低数据密度区域

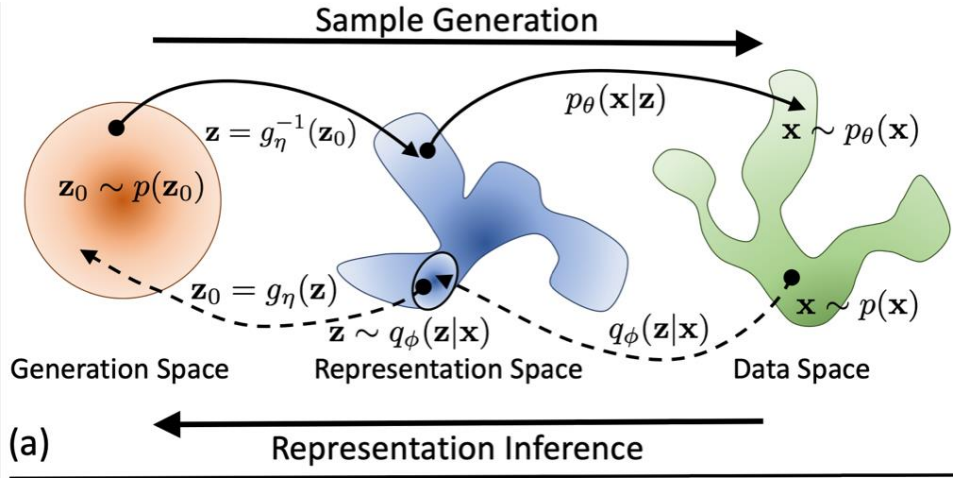


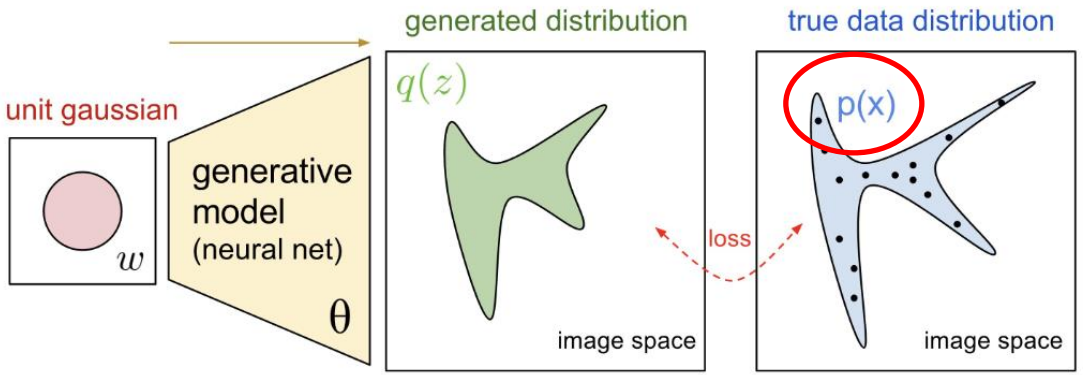
$\nabla_x log p_{data}(x)$

右边： $S_\theta(x)$

算法原理

- 去噪条件打分网络（NCSN）





$S(x) \rightarrow \nabla \log p_\sigma(x)$

- 先通过**DSM**训练一个打分网络$S_\theta(x)$，去估计数据先验的梯度$\nabla_x \log p_{\text{data}}(x)$
- 利用朗之万模拟退火对模型进行采样

算法原理

- 去噪条件打分网络（NCSN）
  - **Prior learning stage先验学习阶段:**

噪声分布为 $p_\sigma(\tilde{x}\mid x)=N(\tilde{x}\mid x,\sigma^2\mathrm{I})$ (1)

则 $\nabla_{\tilde{x}}\log p_\sigma(\tilde{x}\mid x)=-(\tilde{x}-x)/\sigma^2$ (2)

给定一个σ

$$\begin{aligned}\ell(\theta;\sigma) &\triangleq \frac{1}{2}E_{p_\sigma(x)}[\left\|S_\theta(x,\sigma)-\nabla_x\log p_\sigma(x)\right\|_2^2]\\ &=\frac{1}{2}E_{p_\sigma(\tilde{x},x)}[\left\|S_\theta(\tilde{x},\sigma)-\nabla_{\tilde{x}}\log p_\sigma(\tilde{x}\mid x)\right\|_2^2]+C\\ &=\frac{1}{2}E_{p_\sigma(\tilde{x},x)}[\left\|S_\theta(\tilde{x},\sigma)+(\tilde{x}-x)/\sigma^2\right\|_2^2]+C\end{aligned} \quad (3)$$

给定L个噪声，即$\sigma\in\{\sigma_i\}_{i=1}^{L}$

$$L(\theta;\{\sigma_i\}_{i=1}^{I})\overset{\Delta}{=}\frac{1}{I}\sum_{i=1}^{I}\lambda(\sigma_i)\ell(\theta;\sigma_i) \quad (4)$$

算法原理

- 去噪条件打分网络（NCSN）
  - 迭代重建阶段，朗之万动力学模拟退火模型：

Langevin Dynamics:

$$\tilde{x}_t = \tilde{x}_{t-1} + \frac{\alpha_i}{2}\nabla_x \log p_{\sigma_i}(\tilde{x}_{t-1}) + \sqrt{\alpha_i}\, z_t$$

$$= \tilde{x}_{t-1} + \frac{\alpha_i}{2} S_\theta(\tilde{x}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, z_t$$

步长

Algorithm 1 Annealed Langevin dynamics.

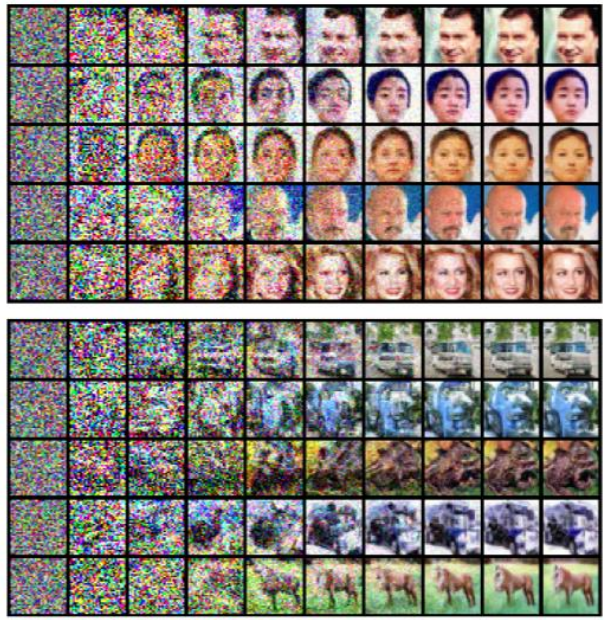
Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\,\mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

实验结果

- 实验结果
  - **IS（Inception Score）**
    - 对于单一的生成图像，**Inception**输出的概率分布熵值应该尽量小
    - 对于生成器生成的一批图像而言，**Inception**输出的平均概率分布熵值应该尽量大
  - **FID分数**
    - **FID分数是在IS基础上的改进**
  - **FID与IS的不同之处在于，IS是直接对生成图像进行评估，指标值越大越好；而FID分数则是通过对比生成图像与真实图像来产生评估分数，计算一个“距离值”，指标值越小越好**

Model	Inception	FID
CIFAR-10 Unconditional		
PixelCNN [59]	4.60	65.93
PixelIQN [42]	5.29	49.46
EBM [12]	6.02	40.58
WGAN-GP [18]	7.86 ± .07	36.4
MoLM [45]	7.90 ± .10	**18.9**
SNGAN [36]	8.22 ± .05	21.7
ProgressiveGAN [25]	8.80 ± .05	-
NCSN (Ours)	**8.87** ± .12	25.32
CIFAR-10 Conditional		
EBM [12]	8.30	37.9
SNGAN [36]	8.60 ± .08	25.5
BigGAN [6]	**9.22**	**14.73**

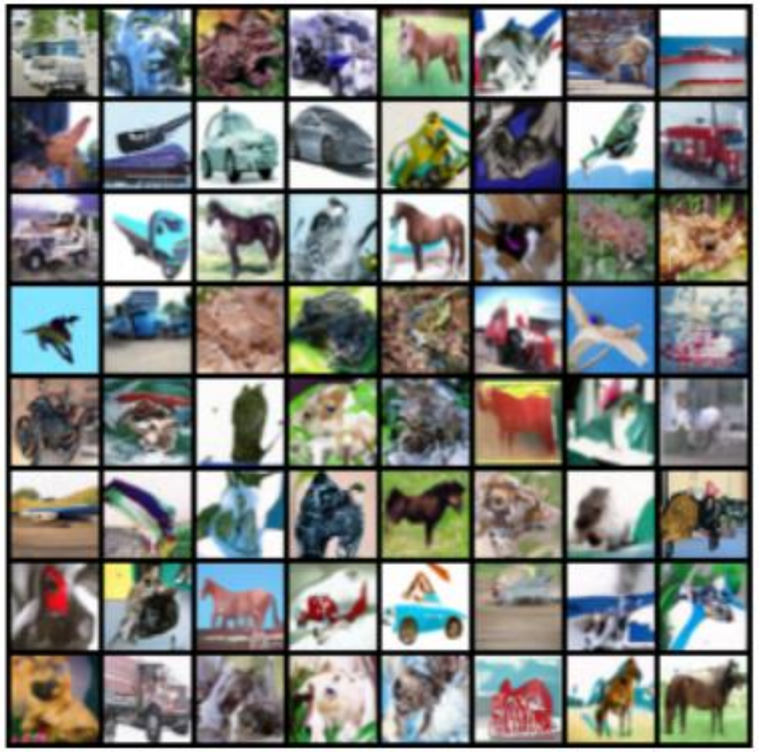


实验结果

- 实验结果-图像生成

(a) MNIST

(b) CelebA



(c) CIFAR-10

实验结果

- 实验结果-图像填充

CelebA (left) and CIFAR-10 (right)

优劣分析

- 横向对比
  - 基于去噪自编码器的生成模型
    - 更具有鲁棒性和泛化性
  - 基于去噪条件打分网络的生成模型
    - 不需要标准化→灵活的框架选择
    - 没有最小化优化→稳定的训练过程
    - 在性能上超过了**GANs** 的效果
- 纵向对比
  - **DAE**算法在当测试样本和训练样本不符合同一分布，即相差较大时，效果不好，明显，**DAE**在这方面的处理有所进步，但有待改善
  - **NCSN**算法中，噪声和模拟退火的噪声等级很关键

应用总结

- 应用领域
  - 回归任务，计算机视觉，语音识别，密度估计，降维，信息获取，自然语言处理
- 未来发展
  - 在语言建模领域的机遇
    - 针对自然语言的大规模深度隐变量模型，该模型使用句子级别的（可变）自动编码器在大型文本语料库进行了预训练， 从而将由符号表达的自然语言组织在一个连续且紧凑的特征空间里，把对句子的语义操作转换为对向量的算术操作
  - 在视觉和语言导航任务上的应用
    - 深度生成模型来合成这些多模态数据，用于预训练，以提高其模型的通用性

参考文献

[1] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol.Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, ICML ’08, pages 1096–1103,New York, NY, USA, 2008. ACM.

[2] Bigdeli, Siavash Arjomand, et al. "Deep mean-shift priors for image restoration." arXiv preprint arXiv:1709.03749 (2017).

[3]Bigdeli, Siavash Arjomand, et al. "Deep mean-shift priors for image restoration." arXiv preprint arXiv:1709.03749 (2017).

[4]Song, Yang, and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution." arXiv preprint arXiv:1907.05600 (2019).

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。